Using TB-Sized Data to Understand Multi-Device Advertising

Completed Research Paper

Quan Wang
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
quanw@andrew.cmu.edu

Beibei Li
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
beibeili@andrew.cmu.edu

Pengyuan Wang
University of Georgia
Athens, Georgia 30602
pengyuan@uga.edu

Jimmy Yang
Yahoo! Inc.
Sunnyvale, California 94089
jianyang@yahoo-inc.com

Abstract

In this study, we combine the conversion funnel theory with statistical model to understand multi-device advertising. We investigate the important question of how the distribution of ads on multiple devices affects the consumer path to purchase. To handle the sheer volume of TB sized impression data, we develop a MapReduce framework to estimate the non-stationary Hidden Markov Model in parallel. To accommodate the iterative nature of the estimation procedure, we leverage the Apache Spark framework on a corporate cloud computing infrastructure. We calibrate the model with hundreds of millions of impressions for 100 advertisers. Our preliminary results show increasing the diversity of device for ads delivery can consistently encourage consumers to become more engaged. In addition, advertiser heterogeneity plays an important role in the variety of the conversion process.

Keywords: Big Data, Cross Device Advertising, Mobile Advertising, Conversion Funnel

Introduction

Cross-device activities are ubiquitous nowadays. People use multiple devices in concert to achieve a common goal. According to Microsoft Research 2013, 57% of consumers use one device to search information related to what they did previously on other devices. And 39% of consumers share content about activities they have accomplished on other devices. Despite the average U.S people spends 60 hours a week consuming content on multiple screens, marketers have limited knowledge about how consumers use devices and what this means for advertisers. For example, cross-device attribution remains an important yet challenging question for both researchers and practitioners. Understanding the cross-device advertising is important whereas ignoring the cross-device effect may lead to overestimating or underestimating the value of a device, hence results in suboptimal resource allocation between devices.

This study is a first step to investigate device interdependence in a non-targeted advertising setting. The research question is how the distribution of ads on multiple devices affects the consumer path to purchase. We look into this question through the lens of a unique large-scale dataset sponsored by a worldwide leading media platform. This dataset provides a full picture of how consumers are moving between devices to interact with advertisers – across desktop, mobile, and tablet – before they convert. In this dataset, we observe the device on which consumers see ads and the device on which conversions subsequently occur. Moreover, we know how the consumers interact with not only the focal advertiser but also all other advertisers on that media platform.

To quantify the marginal impact of a device-specific ad impression on the path to purchase, we follow the conversion funnel theory and construct a non-stationary Hidden Markov Model (HMM). The model allows the marginal impact of an ad to be path dependent and impression specific. After estimating the model parameters, we simulate the path and retrieve the impact of an impression conditional on the path. We estimate the HMM using data from 100 randomly sampled advertisers. The results show that our data contains two types of advertisers. Advertisers of the first type already have engaged consumers at the beginning of the sample period. Their advertising campaigns are also engaging and effective. While advertisers of the second type have dominant unaware consumers initially. The advertising campaigns are not good at engaging consumers either. Interestingly, the two types of advertisers are not visually separable by descriptive statistics, indicating that HMM has revealed some deep patterns about consumer behavior. We also find that there is no optimal device that works best for every advertiser. However, increasing the diversity of device for ads delivery can consistently encourage consumers to become more engaged, everything else being equal.

Methodologically, our paper makes two key contributions. First, we invent a distributed and scalable implementation of the Expectation Maximization (EM) algorithm to estimate HMM. This implementation provides an example of adapting a stand-alone estimation method to the MapReduce manner. The design of the MapReduce framework can be generalized to parallelize other iterative estimation tasks. Second, we calibrate the model on Big Data with the cloud computing techniques. It can produce individual level analysis on hundreds of millions of impressions in less than four hours. This practice has values to marketers, as they probably need to process large scale datasets in their day-to-day work. Moreover, the conversions are rare and sparse compared with impressions. Big Data is important to ensure the statistical power of the estimation.

Several managerial implications can be drawn from our study. To media platform, our analysis shows that device plays an important role in the conversion journal. Ignoring the impact of the device type can result in a misunderstanding of the ads performance. Therefore, the media platform should allow the advertisers to target users by device. Also, it is potentially profitable to price the ads differently on different devices. To advertisers, the heterogeneity shows that not all advertisers are the same. It implies that the successful advertising strategies for other companies may not work universally. It is crucial for the advertiser to understand their own type and take actions accordingly.

The rest of the paper is organized as follows. In the section of literature review, we summarize the current findings o channel interdependence and explain why the device interplay may exist. Then we introduce the unique research context and model free evidence about consumer response to ad exposures. Subsequently, we illustrate how to model consumer funnel using HMM. In the estimation part, we elaborate the parallelized implementation details using Spark and corporate cloud. Then we show the estimation results of many advertisers and one typical advertiser. The last section concludes the study and discusses our limitations.

Literature Review

Our paper builds upon three different streams of literature: (i) channel interdependence in advertising, (ii) Hidden Markov Models (HMM) in computer science and marketing, and (iii) the value of Big Data.

Channel Interdependence in Advertising

Due to the growing importance, researchers have an increasing interest in the interdependence between advertising channels. There are three types of channel interdependent relationships: complementary, substitutive, and independent (Goldfarb and Tucker 2011). Complementary relationship means the presence of one channel can increase the value of another channel. This relationship is also called as spillover effect (e.g Li and Kannan 2014) and synergies (e.g Yang and Chose 2010, Naik and Raman 2003). Substitutive relationship, on the other hand, means showing an advertisement on both channels performs the same as or even worse than showing only on one channel. This relationship is also called as cannibalization effect (e.g. Gopal et al. 2011). And independent relationship means the channels operate separately. Therefore, the activities on one channel do not affect the activities on another channel. It is worth mentioning that the interdependent relationship can be asymmetric, i.e. the impact of one channel on the other channel can be complementary, whereas the impact vice versa can be substitutive. For

instance, Ghose et al. 2013 found that using the online and smartphone advertising channels simultaneously improves web conversion rates but decreases mobile conversion rates. In the online shopping scenario, Kaiquan et al. (2015) found the introduction of tablets as a new shopping channel can increase the sales volume and revenue of e-commerce platforms, suggesting that the tablet channel acts as a substitute for the PC channel while it acts as a complement for the smartphone channel.

There are several plausible reasons suggesting why the channel interdependence exists among the tablet, smartphone, and desktop devices. Compared with desktop, mobile devices provide portability, mobility and ease of access to touch screens and device tailored applications (Adipat et al. 2011, Gerpott et al. 2013). Therefore it is plausible that the micro-moments from one device could enhance the impact of an advertising exposure on another device. On the other hand, it is also possible that the devices may not have a significant impact on the conversion outcomes. If the content of the ads is the only determinant of the purchase decision, the device type will have no impact on conversion. Therefore, the overall effect of advertising devices on advertising outcomes remains an empirical question.

Hidden Markov Models and Conversion Funnel Theory

Our paper follows the methodology of Hidden Markov Models. A Hidden Markov Model (HMM) is a type of Bayesian network that models sequential data. In this model, the data is assumed to be a Markov process with unobserved (hidden) states. Specifically, the underlying states are assumed to follow a Markov chain, which means given the present, the future state is independent of the past states. Conditional on the present state, the outcome is independent of other outcomes and states. While the state sequence is not directly visible, the outcome sequence, dependent on the state, is visible. HMMs are especially known for their successful applications in speech and handwriting recognition, part-of-speech tagging, and bioinformatics. Recently, HMMs have been introduced to model consumer behavior in various business and managerial contexts. For example, Netzer et al. (2008) model customer-brand marketing interactions to relate the underlying customer relationship with observed buying behavior. Hauser et al. (2009) infer the underlying cognitive styles of users from observed clickstream data. Singh et al. (2011) identify developers' underlying learning dynamics from their experience and interaction with peers in the context of open source software development.

In the context of digital advertising, HMM has good potential because the model structure is coherent with an important conceptual framework, conversion funnel theory, which describes the consumer navigation journey. The theory points out that a consumer experiences several psychological stages to reach a purchase decision. The conversion funnel framework can be represented by HMM where the psychological stages correspond to the underlying states, and the purchase decision corresponds to the observed outcome. Abhishek et al. (2012) are among the first researchers to use HMM to model advertising attribution. We follow their paper and also choose HMM as our main methodology. Our current study has not shown whether the HMM model fits the data better than alternative models. In the future work, we will provide a comprehensive model comparison.

Big Data

Big Data is used to describe the data sets and analytical techniques in applications that are so large and complex that they require advanced data storage, management, analysis, and visualization technologies (Chen et al. 2012). There are four aspects of Big Data: velocity, volume, variety, and veracity. Big Data represents a new era in data exploration and utilization (Zikopoulos et al. 2011). It has the potential to lead disruptive changes (Baesens et al. 2014) in various high-impact domains such as e-commerce, market intelligence, e-government, healthcare, and security.

In terms of techniques, distributed systems and open source techniques have enabled the rapid progress on big data. For example, Kafka and Kinesis are distributed messaging systems that collect high volumes of log data with low latency. NoSQL databases, such as HBase, MongoDB, DynamoDB, and Cassandra, together with Hadoop Distributed File System (HDFS), provide storage solutions that go beyond the tabular relations used in relational databases. Open source framework such as MapReduce, Spark, Hive, and Pig support the analysis of large and diverse datasets across clustered systems. And data visualization tools such as D3.js and Tableau help the analysts understand and present the result fast and clearly.

In terms of statistical theory and application, there are still challenges in developing the theoretical principles needed to scale inference and algorithms to massive scale (Jordan and Mitchell 2015). One natural way to deal with the challenges of big data is to make the data smaller. That is, to seek a compact representation of the data so that certain properties are preserved. For example, sketches find carefully designed random projections of the input data but reduces the high dimensions. Conceptually, these studies aim at finding sufficient statistics for properties of the data. In the field of digital marketing, researchers have adopted advanced structural models to better explain consumer behavior. However, to the best of our knowledge, few structural models are incorporated with Big Data. This study leverages the distributed and parallel computing techniques to explore the value of massive data.

Research Context and Data Description

Research Context

Today advertisers have a vast array of choices to communicate with the consumers. This study focuses on an emerging and promising type of ads: native ads. Native ads adapt the advertising information to the surrounding media organic content in a natural way. In contrast with traditional banner ads that fight with media content for space and attention, native ads match the visual design, format, and display location of the organic content. Meanwhile, they are visually distinguishable from the organic media content as they are labeled as "sponsored" at the top of the ads. Nowadays native ads are pervasive on digital media platforms such as Facebook News Feed, Twitter tweets, Yahoo news, New York Times, et al. Native ads are apt to drive much better performance than non-native experiences by providing a less disruptive experience (Business Insider 2015). They are especially important on mobile devices because mobile screens cannot seamlessly present a banner ad without being intrusive (Venture Beat 2015). Native ads have grown tremendously since 2012. Marketers expect that US native ad spending would continuously raise double-digit growth, reaching 8.8 billion dollars in 2018 (eMarket 2014).

Our data is obtained from a major digital media company (for confidential reasons, it's called Company X onwards) that has worldwide leading media products and advertising solutions. Our research setting is unique in a few aspects, which help us answer the research question from the causal perspective. First, the user is tracked through a combination of login, probabilistic identity matching, and cookie-based technology. The data enable us to track the user trajectory on PC, tablet, and mobile devices. Second, for any U.S. user of the Company X, the data contains all her native ads impressions for 22 days between April and May 2015. The data is not limited to one single advertiser or a specific industry. Instead, it covers impressions of all advertisers working with Company X's native ads business. For each impression, we know detailed information about the advertiser id, advertiser name, whether it is clicked or not, whether it leads to a conversion or not, the device type of the display, the media context of the display, campaign id, consumer id, gender, age, date and time. Third, complex targeting strategies such as contextual targeting, behavior targeting, and retargeting were not enabled in the campaigns of our sample. The impressions were almost randomly displayed after controlling for the consumer age, gender, and geo-location. Unfortunately, the dataset does not include the user exposure to other types of marketing mix such as paid search ads, banner ads, email promotions, social media, etc. We also discuss the data limitation in the conclusion section.

Consumer Level Summary Statistics

Here are some summary statistics conducted at the consumer level. In our sample, there are over 146 million consumers that have been exposed to native ads by over 6,000 advertisers, among which 25 million users had at least one click incidence and 361 thousand users had at least one conversion incidence during the sample period. In total, all the consumers were exposed to tens of billion impressions. They performed about millions of clicks and hundreds of thousands of conversions. Table 1[1] reports the summary statistics of the consumer level variables. The median consumer was exposed to 17 impressions from 8 advertisers during our sample period. But the click and conversion activities were

[1] The reported dataset and results are deliberately incomplete and subject to anonymization, and thus do not necessarily reflect the real portfolio at any particular time.

much less frequent. There is significant heterogeneity in consumers' ad impression, click, and conversion frequencies. The maximum impression frequency and the maximum click frequency are abnormally high. To exclude possible manipulated and malformed data, we remove the consumers with 99.85 percentile of impressions and clicks in our final analysis.

Table 1. Consumer Level Summary Statistics

	Mean	Sd	Median	Min	Max
Impression Frequency	≈110	283.567	17	1	>130,000
Click Frequency	0.410	1.661	0	0	>1,000
Conversion Frequency	0.005	0.529	0	0	29
Unique Advertisers	≈30	52.160	8	1	>1,000
Age	34.739	15.636	31	12	78
Female	0.536	0.462	1	0	1

Do consumers respond to advertisement differently given their different devices in use? To have an exploratory understanding of this question, we plot the ad response by device usage segment in Figure 1. Device usage is approximated by which devices the consumer had been observed to use in our impression data. This proxy is reasonably good because the consumer is inevitably exposed to native ads whenever she browses the Company X's media platform, hence tracked by our dataset. The more devices a consumer use, the more Internet savvy she is. In Figure 1, the light gray bar represents the mean click-through rate of consumers falling into the corresponding segment of device usage. The dark gray bar represents the mean ratio of conversion divided by click for that segment. The error bar represents 95% confidence interval of the mean. The figure shows that consumers using a single device are more likely to click an ad than those consumers using multiple devices. However, once have clicked the ad, consumers using a single device are typically less likely to convert than consumers using multiple devices. Unfortunately, this exploratory analysis cannot be interpreted as causal. It remains unclear whether the device usage leads to different ad response patterns, or whether people responding to ads differently happen to have different device usage.

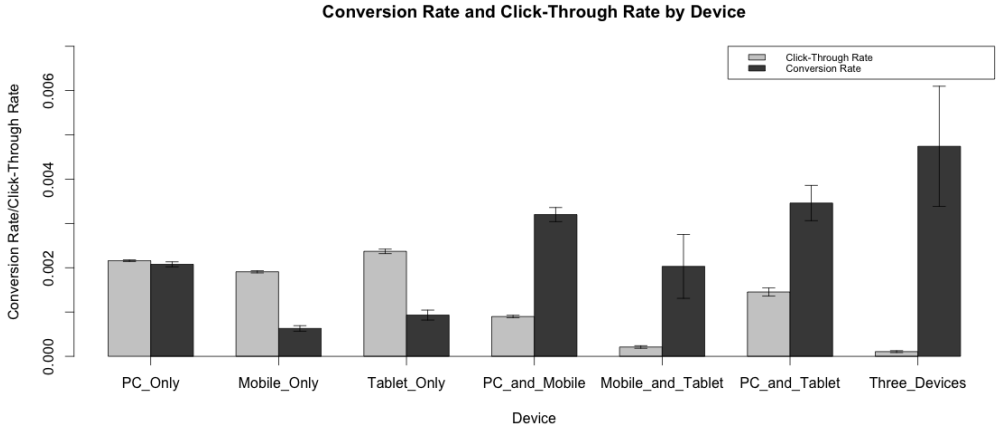

Figure 1. Conversion Rate and Click-Through Rate by Device

Advertiser Level Summary Statistics

Now let's zoom in to take a look at the advertisers. Table 2[2] summarizes the main variables at the advertiser level. The summary statistics show that there is a good amount of heterogeneity among advertisers. The median advertiser has acquired a few millions of impressions, thousands of clicks, and 45 conversions from about 180,000 users for their native ads campaigns. The median click-through rate is

[2] The reported dataset and results are deliberately incomplete and subject to anonymization, and thus do not necessarily reflect the real portfolio at any particular time.

0.3% and the median conversion rate is 0.63%. However, the maximum impressions, clicks, and conversions can be hundreds of times higher than the median values. In contrast, the minimum occasions are very few for small advertisers. Therefore we remove the advertisers that have less than 30 conversion occasions because we are concerned about the statistical power of few conversions. Among the remaining advertisers, 92.4% advertisers have impressions on all of the three devices, 2.1% advertisers have impressions on one device only, and 3.4% advertisers have impressions on two devices. We keep the 92.4% advertisers in our final sample. Dropping advertisers from the raw dataset would not introduce sampling bias, as the allocation of impressions on the device is randomly generated by the media platform instead of strategically decided by advertisers.

Table 2. Summary Statistics By Advertiser

	Mean	Sd	Median	Min	Max
Impression Frequency	≈14,000,000	20,139,364	≈3,000,000	1	>400,000,000
Click Frequency	≈50,000	204,286	≈9,000	0	>3,000,000
Conversion Frequency	≈900	3,092	45	0	≈ 30,000
Unique Consumers	≈700,000	1,794,283	≈200,000	1	>2,000,000
Click-Through Rate	0.0041	0.0038	0.0030	0	0.0421
Conversion Rate	0.0342	0.0996	0.0063	0	0.3371

Do the ads perform differently for different advertisers? We take a glance at this question by plotting the click-through rate and the conversion rate in Figure 2. In this figure, each dot represents an advertiser. The size of the dot is proportional to the total number of impressions. The x-axis and y-axis are logarithm transformed conversion rate and logarithm transformed click-through rate. The figure shows that regardless of how many impressions have been delivered, advertisers are very heterogeneous in the ad performance. For the advertisers on the upper right corner of the figure, they have good performance on both metrics, indicating that they have used the native ads and their own website wisely. For advertisers on the upper left corner of the figure, they have relatively high click-through rate but low conversion rate, implying that their ads campaigns are attractive but their own website cannot retain many consumers. For the advertisers on the lower right corner, they have high conversion rate but low click-through rate, meaning that they are good at retaining the very engaged consumers. But they need to improve the advertising campaigns to attract unaware and potential consumers. Finally, for the advertisers on the lower left corner, they perform badly at both metrics.

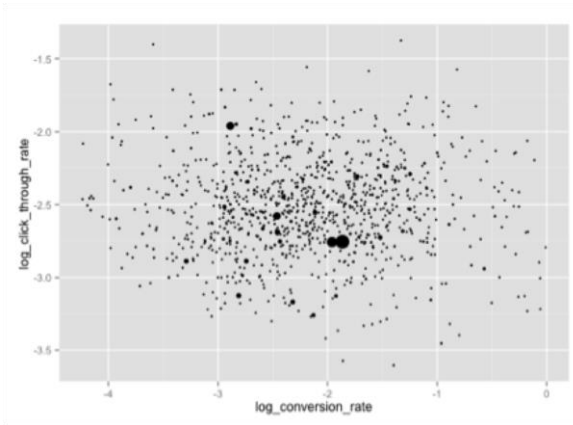

Figure 2. Conversion Rate vs. CTR

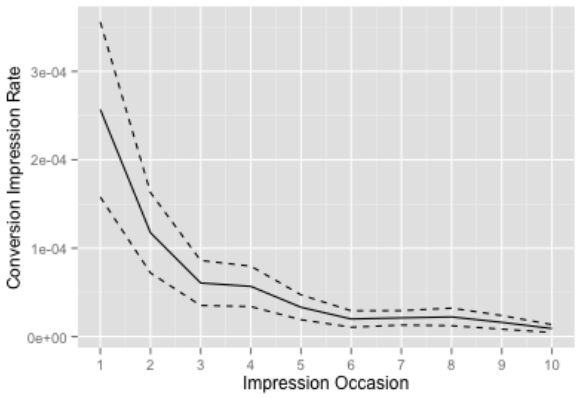

Figure 3. Decreasing Time Trend of Conversion

Model Free Evidence

This section presents some patterns in the data that suggest how consumers respond to ads in different circumstances. We show that the diversity of device drives consumers to convert more quickly. Consumers can be annoyed by the repeated display of the same ads. The model free evidence also highlights the variation of the data at the advertiser level.

Device Diversity

First of all, we argue that a consumer may convert more quickly if the exposure incidences are across devices rather than on the same device. This is because people tend to use different devices in different places and contexts. The same ads displayed on multiple contexts could remind the user of the previous impressions. It can also potentially engage the consumers more effectively because of the familiarity. If that is the case, we should observe a positive correlation between the device diversity and the conversion likelihood. Therefore we test the hypothesis with the following model specification. We assume the conversion can be modeled as

$$Conversion_{ij} = b_1 DeviceDiversity_{ij} + b_2 Click_{ij} + u_i + v_j + e_{ij}$$

where Conversion_{ij} is whether consumer i has ultimately converted for advertiser j during the sample period. $\text{DeviceDiversity}_{ij}$ is the number of devices that consumer i has used to receive native ads about advertiser j. Click_{ij} is the number of clicks the consumer i has conducted about advertiser j. u_i is the individual fixed effect and v_j is the advertiser fixed effect. We run this regression for 100 randomly sampled advertisers and all the exposed consumers. The results are reported in Table 3. The positive and statistically significant coefficient β_1 indicates that a consumer is more likely to buy if she sees the ads on multiple devices, compared with the other case that she sees the ad with the same frequency but on one single device. Notice that the value of β_1 is close to zero. However, it does not indicate the effect is trivial. The magnitude is driven by the common small conversion rate.

Table 3. Device Diversity and Conversion

Model Coefficient	Estimation
b_1	0.00105*** (0.00037)
b_2	0.00055*** (0.00004)
User Fixed Effect	Yes
Advertiser Fixed Effect	Yes
R2	0.0449

Displeasure of Repeated Ads

Next, we explain the temporary change in the effect of ad exposures. The idea is the marginal effect of an ad on conversion is not always constant. Instead, it can vary by how many times the ad has already been displayed to the consumer. To explore the temporary pattern, we plot the conversion impression rate by the impression occasion in Figure 3. In this analysis, if the impression occasion of an ad is 3, it means this is the 3rd ad impression received by the focal consumer about the focal advertiser. We use the previously randomly sampled 100 advertisers. We calculate the conversion impression rate for each of the advertisers and plot the average rate and 95% confidence interval of the average rate. Figure 3 shows a clear downward trend, meaning that the conversion is more likely to happen in the early impression occasions than in the repeated occasions. This graph indicates that repeated ads might annoy consumers.

Advertiser Heterogeneity

The model free evidence also indicates that advertisers play an important role in the variety of the data. To show the advertiser heterogeneity, we apply a simple logistic regression to each of 100 randomly sampled advertisers. Then we summarize the coefficients using the histogram. If the advertiser level heterogeneity is minimal, we should expect the vast majority of the mass to be concentrated at a few values far away from zero. The logistic regression model is as follows.

$$Conversion_{it} = MobileIndicator_{it} + TabletIndicator_{it} + DeviceStock_{it} + Homepage_{it} + Mail_{it} + InpressionStock_{it} + ClickStock_{it} + u_i + e_{it}$$

The distribution of the model coefficients is reported in Figure 4. Most of the coefficients have the mass density at zero, indicating that advertiser level heterogeneity might be an important source of data variation. In contrast with the distributions of other coefficients that can be either positive or negative, the coefficient of device stock is consistently positive at the range of [0.414, 0.998]. It shows from another angle that device diversity is positively associated with fast conversion.

It is also worth mentioning that all the coefficients of the logistic regression models are statistically significant. The standard errors are in the magnitude of 10^{-5}. So all the coefficients are precisely estimated. And most of the coefficients are significantly different from zero. Figure 4 is not saying the device of display, the context of display, or consumer historical behavior are not a significant determinant of conversions. Instead, it highlights the heterogeneity among advertisers.

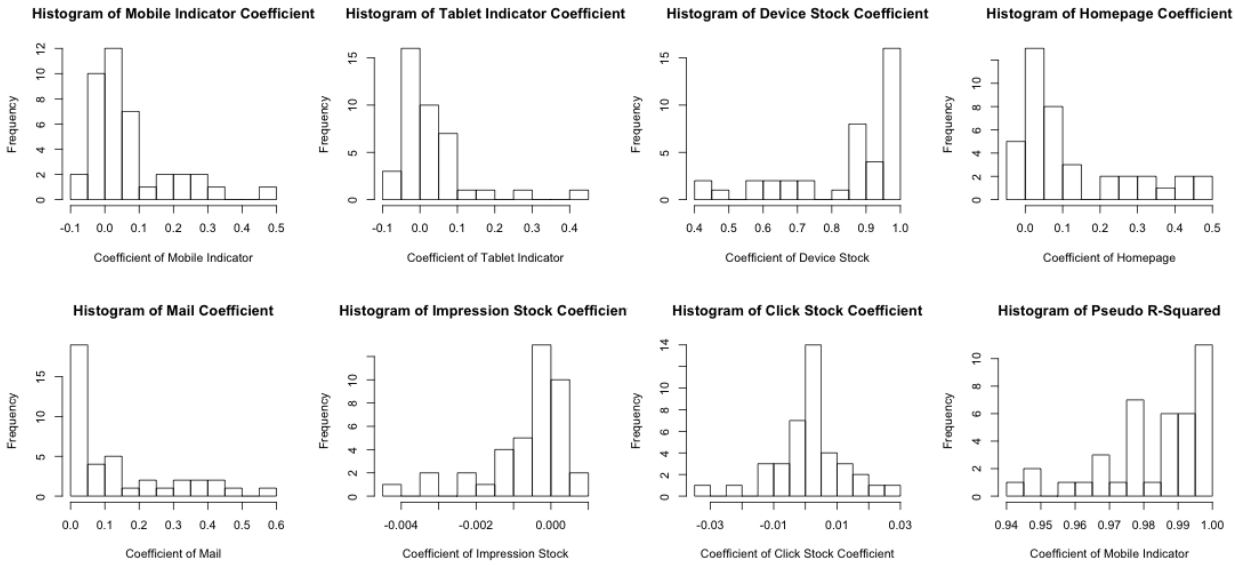

Figure 4. Coefficients of Logistic Regression

Model

Conversion Funnel Theory

The conversion funnel theory provides a theoretical framework to model the deliberation process of consumers prior to purchase. The literature on conversion funnel theory suggests that new consumers are unlikely to buy immediately in the first visit to an advertiser's website. Instead, it takes time for consumers to discover the company, evaluate the brand and product, determine the fit, and make the purchase decision. Therefore there are a few stages that consumers would experience on the path to purchase. Compared with those in late stages, consumers in early stages are less engaged and less likely to buy. Although the engagement level is unobserved typically, it can be inferred from the sequence of purchase decisions.

In our model, the ads fit in the conversion funnel framework by affecting when and how consumers move between stages. It is assumed that effective ads encourage consumers to move from less engaged stages to more engaged stages, whereas ineffective ads do not lead to stage movement, or even discourage the movement to more engaged stages. Also, for advertising channels that have complementary effects, showing ads on both channels should encourage consumers to be more engaged. In contrast, showing ads on substitutive channels may encourage consumers to move back to less engaged stages. The timing of the model is as follows. At each moment, the consumer receives native ads, if there is any. Due to the impact of the ad, she becomes more or less engaged with the advertiser, or stays the same. She decides whether to click the ad or buy the product. Her click and conversion decisions are observed and next moment comes.

Model Specification

In accordance with the conversion funnel theory, we construct a Hidden Markov Model (HMM) with three vector components: the observed conversion decisions: $C_i = \{C_{i1}, C_{i2} \dots C_{iT_i}\}$, the observed advertising exposures $a_i = \{a_{i1}, a_{i2} \dots a_{iT_i}\}$, and the unobserved engagement states $S_i = \{S_{i1}, S_{i2} \dots S_{iT_i}\}$ for consumer i. The time indicators 1 to T_i represent the exposure occasions of the consumer on the conversion path. At any time t, the consumer's conversion result $C_{it} \in \{0,1\}$ is a binary random variable. The distribution of this random variable, which represents the propensity to convert, is assumed Bernoulli and only depends on the current hidden state S_{it}. The transitions between different states are assumed to follow a Markov process, i.e. the transitions out and into a particular state depend only on the current state and not on the previous path that the consumer took to get into the state. The Markov process can be non-stationary, i.e. the transition probabilities are not always the same for any context. Instead, they can be affected by the impression occasions. Consumer i's exposure to advertisement a_{it} is a column vector of covariates that captures the current ad exposure as well as the cumulative ads stock until t. The hidden state $S_{it} \in \{S_1, S_2 \dots S_{|S|}\}$ is a stochastic variable that belongs to a discrete state space. The hidden state is unobserved but can be inferred from the observed conversion outcomes and ad impression information.

In all variants of HMM models, there are three groups of model primitives corresponding to the transition matrix, conversion matrix, and initial distribution. Below, we describe each of the primitives in our model specification.

We use the transition matrix Q_{it}, a square matrix with size $|S|$ by $|S|$, to represent the transition rule between hidden states. The element q_{itjk} in the jth row and kth column of Q_{it} denotes the transition probability from state j at t to state k at t+1 for consumer i. The functional form of the transition probability is assumed as

$$q_{itjk} = p(S_{it} = S_k | S_{it-1} = S_j) = \frac{\exp(\rho_{jk} a'_{it} + e_{itjk})}{\sum_{j=1}^{|S|} \exp(\rho_{jl} a'_{it} + e_{itjl})}$$

where ρ_{jk} is the effectiveness parameter that captures how the advertising characteristics affect the consumer's propensity to transition from state j to state k. ρ_{jk} is assumed to vary across states but is shared by individual consumers and is constant over time. The random shock in the transition probability is captured by the unobserved error term

$$e_{itjk} \sim N(0, \sigma^2)$$

Specifically, a_{it} represents the determinants of the transition probability that capture the ad related attributes the consumer is exposed to. Following existing work on attribution modeling and advertisement effectiveness, we define a_{it} as a vector of advertising characteristics that is $\{\text{DeviceIndicator}_{it}, \text{DisplayContextIndicator}_{it}, \text{DeviceStock}_{it}, \text{ImpressionStock}_{it}, \text{ClickStock}_{it}\}$. The transition determinants a_{it} also include an intercept constant of 1.

At any time, the consumer's conversion decision only depends on her current state. We model the conversion probability as a state specific constant that is the shared by all individual consumers. In particular, let B_{jk} denote the probability of observing the conversion outcome k given the state S_{it}. Following the related literature, we assume the probability of converting increases as the consumers move down the conversion funnel and become more engaged.

$$B_{jk} = p(C_{it} = k | S_{it} = S_j)$$

The initial state membership is the conversion probability of the consumer in the absence of any advertisement. Specifically, let π_{ik} denotes the probability that consumer i is in state k with the absence of ads incidence. The functional form of the initial membership is specified as

$$\pi_{ij} = p(S_{i0} = S_j) = \frac{\exp(\phi_j z'_i + \varepsilon_{ij})}{\sum_{l=1}^{|S|} \exp(\phi_l z'_i + \varepsilon_{il})}$$

where z_i represents the determinants of the initial probability of consumer i. ϕ_j is the state specific coefficient vector that captures how the user demographics affect the consumers' initial states. The unobserved random shock that affects consumer initial distribution is assumed to follow a normal distribution

$$\varepsilon_{ij} \sim N(0, \upsilon^2)$$

We assume the initial probability is affected by the determinants of the initial probability z_i as follows:

$\{DeviceDiversity_i, Advertiser_i, TotalImpressions_i, TotalClicks_i, TotalConversions_i, Age_i, Gender_i\}$. Same as in a_{it} the initial determinants z_i also include an intercept constant of 1. Let $\Theta = \{\rho, \phi, B, \sigma^2, \upsilon^2\}$ denotes the set of parameters of our model. The next section provides the outline of our estimation strategy to recover the parameters.

Estimation

For smaller sized samples that typically include tens of thousands of consumers, researchers have several choices about analytical tools and estimation strategy. However, when dealing with large-scale datasets, many tools and algorithms are no longer applicable. There are at least two challenges in Big Data problems. First, the sheer volume of data requires a large amount of processing and storage resources. The problem is solvable only with supercomputers or cloud computing service. Second, when the estimation procedure is iterative, the communication cost between distributed computing workers becomes very expensive.

Distributed Implementation

Our project takes an original approach to estimate the non-stationary HMM model. We adapt the estimation framework to the MapReduce manner. We also use an emerging Big Data paradigm, Apache Spark, to run the estimation on the cloud. The core innovation of our estimation procedure is to pick the right estimation algorithm, partition the raw data wisely and use sufficient statistics in a nice way. Our approach consists of two stages, the map stage and the reduce stage. In the map stage, the computational workers take the global model parameters and the distributed data as input. They compute the sufficient statistics for each data partition separately. In the reduce stage, the sufficient statistics are gathered together and used to update the global parameters. The procedure stops when the global parameters converge.

In particular, we recover the Maximum Likelihood Estimator (MLE) by maximizing an expected-complete-log-likelihood function using the Expectation Maximization (EM) method (e.g. Rabiner 1989, Sahoo et al. 2012). The EM algorithm is an efficient iterative procedure to solve statistical estimation problems. It computes the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. Each iteration of the EM algorithm consists of two processes: the E-step and the M-step. In the E-step, the missing data are estimated given the observed data and the current estimate of the model parameter. This is achieved using the conditional expectation. In the M-step, the likelihood function is maximized under the assumption that the missing data are known. Convergence is assured since the algorithm satisfies the Jensen's inequality and is guaranteed to increase the likelihood at each iteration. Compare with the alternative Markov Chain Monte Carlo (MCMC) approach, the EM approach converges faster (Ryden 2008). The complexity of the EM method is linear to the number of impressions in the data, whereas the complexity of directly solving the MLE is exponential to the number of impressions.

Denote the total number of consumers are N. Assume their ads impressions and conversions are independent. The total probability of observing the data is $p(C_1, C_2 \dots C_N; \Theta) = \prod_{i=1}^{N} p(C_i; \Theta)$. Define $Q(S_i; \Theta) := p(S_i|C_i; \Theta)$. Then the EM estimator is the solution of the following optimization problem with linear constraints.

$$\widehat{\Theta} := arg \max_{\Theta} \prod_{i=1}^{N} p(C_i; \Theta) = arg \max_{\Theta} \sum_{i=1}^{N} \sum_{S_{it}=S_1}^{S_{|S|}} Q(S_i) \log \left[\sum_{m=1}^{|S|} \pi_{im} \prod_{t=1}^{T_i} B_{S_{it} C_{it}} \, q_{it S_{it-1} S_{it}} \right]$$

$$s.t. \sum_{m=1}^{|S|} \pi_{im} = 1, \pi_{im} \in [0,1];\ \forall i = 1,2,\dots,N$$

$$\sum_{C_{it}=0}^{1} B_{S_{it}C_{it}} = 1, B_{S_{it}C_{it}} \geq 0;\ \forall i = 1,2,\dots,N, t = 1,2,\dots,T_i$$

$$\sum_{S_{it}=S_1}^{S_{|S|}} q_{itS_{it-1}S_{it}} = 1, q_{itS_{it-1}S_{it}} \geq 0;\ \forall i = 1,2,\dots,N, t = 1,2,\dots,T_i$$

To solve this optimization problem, we construct the Lagrangian multiplier and set the partial derivatives of the Lagrangian functions to zero. Then we can represent $\Theta = \{\hat{\rho}, \hat{\phi}, \hat{B}, \widehat{\sigma^2}, \widehat{\upsilon^2}\}$ as functions of four sets of probabilities $\xi_{itjk} \coloneqq p(S_{it} = S_j, S_{it+1} = S_k | C_i; \Theta)$, $\gamma_{itj} \coloneqq p(S_{it} = S_j | C_i; \Theta)$, $\alpha_{itj} \coloneqq p(C_{i1}, C_{i2}, \dots C_{it}, S_{it} = S_j; \Theta)$, and $\beta_{itj} \coloneqq p(C_{it+1}, C_{it+2}, \dots C_{iT_i} | S_{it} = S_j; \Theta)$. These four variables are the sufficient statistics of the individual conversions C_i and the globally shared parameters $\hat{\Theta}$. Therefore, we partition the data by individual consumers. In the map stage, the individual consumers' raw data and the $\hat{\Theta}$ of the previous iteration are used to calculate sufficient statistics $\{\xi_{itjk}, \gamma_{itj}, \alpha_{itj}, \beta_{itj}\}$ in parallel. In the reduce stage, the $\hat{\Theta}$ of the current iteration are updated by combining the individual $\{\xi_{itjk}, \gamma_{itj}, \alpha_{itj}, \beta_{itj}\}$. Due to the limit of space, we do not include the technical details of the updating process in this paper. They are available upon request.

Rescaling and Confidence Interval

Rescaling is usually needed for the EM algorithm. This is primary because α_{itj} and β_{itj} are probabilities associated with very specific conversion sequences and very specific hidden states. These probabilities go to zero very quickly as t becomes sufficiently large. To ensure the robustness of the implementation, we rescale α_{itj} by multiplying it by the scaling factor $w_{it} = \frac{1}{\sum_{j=1}^{|S|} \alpha_{itj}}$. We rescale β_{itj} by dividing it by w_{it}. After rescaling the sufficient statistics, the model can appropriately handle very skewed dataset. For example, if most of the consumers receive several impressions but a few consumers receive tens of impressions for some reason, the model can still estimate the coefficients without incurring any machine errors. In addition to rescaling the intermediate statistics in the model, we rescale the data a_{it} and z_i such that each dimension of a_{it} and z_i is centered at zero with a standard deviation of 1. This step may not be necessary for smaller scaled HMM problems. But when the data is large and heterogeneous at both consumer level and advertiser level, rescaling the data allows the parameters to converge to stable maxima with any arbitrary starting point.

When HMM model and EM algorithm are discussed in the machine learning and data mining domain, the confidence intervals of the parameters are often understated. However, in our setting, we need to understand how precise is the estimated treatment effect. The confidence intervals of the parameters are important to us. For the conversion probabilities B, we use the Bootstrap method to get the confidence interval. And for the initial and transitional coefficients $\{\hat{\rho}, \hat{\phi}\}$, the confidence intervals have closed-form analytical solution. We calculate the confidence interval for the slope of the following OLS regressions:

$$f_j \begin{bmatrix} z_1' \\ z_2' \\ \dots \\ z_N' \end{bmatrix} + \begin{bmatrix} e_{1j} \\ e_{2j} \\ \dots \\ e_{Nj} \end{bmatrix} = \begin{bmatrix} \log p_{1j} \\ \log p_{2j} \\ \dots \\ \log p_{Nj} \end{bmatrix}, 1 \leq j \leq |S| \qquad r_{jk} \begin{bmatrix} a_{11}' \\ a_{21}' \\ \dots \\ a_{N1}' \\ a_{12}' \\ a_{22}' \\ \dots \\ a_{NT_N}' \end{bmatrix} + \begin{bmatrix} e_{11jk} \\ e_{21jk} \\ \dots \\ e_{N1jk} \\ e_{12jk} \\ e_{22jk} \\ \dots \\ e_{NT_Njk} \end{bmatrix} = \begin{bmatrix} \log q_{11jk} \\ \log q_{21jk} \\ \dots \\ \log q_{N1jk} \\ \log q_{12jk} \\ \log q_{22jk} \\ \dots \\ \log q_{NT_Njk} \end{bmatrix}, 1 \leq j \leq |S|$$

Identification

There are a few identification issues to handle in the proposed HMM model. We first discuss how our model parameters $\hat{\rho}, \hat{\phi}$ are identified. They are the state specific fixed-effect parameters that are constant across consumers. These parameters are easily identified if there is sufficient variation in the demographics, ad exposures, and final conversions (Abhishek et al. 2012). The consumer and occasion specific unobserved heterogeneity ε, ϵ are not directly observed. They capture the idiosyncratic factors impacting the initial state and transition behavior but is unknown to the researcher. We assume they are drawn from the same normal distribution. Then we can estimate the variance of the normal distribution using the OLS regressions.

When estimating the model parameters $\hat{\phi}$, we fix ϕ_j for one random $j \in \{1 \dots |S|\}$ to be zero. This is because the transition probability across states sums up to 1. One degree of freedom is lost for that. Similarly, when estimating $\hat{\rho}$, we fix $\rho_{j1}, \rho_{j2} \dots \rho_{j|S|}$ for one random $j \in \{1 \dots |S|\}$ to be zero because the transition probabilities from state S_j to states S_1 to $S_{|S|}$ sum up to 1. After fixing one dimension to be zero, the model can identify the rest of the transition parameters.

In an HMM, we need to identify both the model parameters and the hidden states. HMM typically suffers from the labeling switching problem, i.e. the states are not explicitly labeled by the engagement level. We address this issue by enforcing the conversion probabilities $B_{i1,1} \leq B_{i2,1} \leq B_{i3,1}$, i.e. consumers are more likely to convert as they move down the conversion funnel.

Estimation Performance

We report the computational performance of the estimation method on two computational infrastructures in Table 4. The first infrastructure is a personal computer with a dual-core Intel Core i7 CPU and 16GB onboard memory. The running program is a piece of stand-alone R code implementing the EM algorithm. The second infrastructure is a corporate cloud with sufficient CPU Cores and hundreds of TB sized memory. The running program is the proposed distributed EM algorithm implemented in PySpark. Our experiment shows that the runtime of the PySpark program is nonlinear to the size of the data, primarily because it takes a fixed amount of time for the Yarn resource manager to spawn a full-fledged program no matter how large or small the problem is. But the Spark program is very easy to scale up. However, for the program running on the personal computer, the runtime is linear to the data size. When the size of the data exceeds millions of records, the stand-alone program becomes very slow. And it cannot handle the full size of the sample due to memory constraints.

Table 4. Estimation Time in Minutes

Size of Data (Impressions)	Corporate Cloud	Personal Computer
1,000	12	0.4
10,000	14	6
100,000	17	51
1,000,000	31	376
> 50,000,000	~180	NA

Empirical Analysis

To apply the proposed model to our dataset, we randomly sample 100 partitions corresponding to 100 advertisers. We estimate the generic HMM on each of these advertisers. The sample sizes of the advertisers are summarized in Table 5.

Table 5. Sample Size of 100 Advertisers

	Mean	S.D.	Median	Min	Max
Impression Frequency	27,774,731	59,159,455	7,058,810	309,279	451,042,369

Estimation Results of 100 Advertisers

This subsection introduces the overall estimation results of these advertisers. Unfortunately, it is not helpful to report the raw estimated parameters $\widehat{\Theta}$ because of the labeling switching problem. To compare advertisers in a meaningful way, we conduct a few simulations to transform the estimation results to probabilities. In the transformation of initial probabilities, we generate 120 hypothetical consumers. The demographical features of these consumers are representative of the population. The initial probability of an advertiser is approximated by the average value of the 120 initial probabilities corresponding to these hypothetical consumers. Similarly, in the transformation of transition probabilities, we simulate multiple impression occasions. Specifically, we generate 324 sequences of ad impressions that cover all combinations of device usage for a sequence of 4 ads impressions. Then we acquire the mean transition probabilities as the average of the transition probabilities in all 324 scenarios. Finally, for the conversion probabilities, no simulation is needed because they are estimated directly from the model.

After acquiring the probabilities, we cluster the advertisers using the hierarchical clustering method. The dendrogram of the hierarchical clustering analysis shows that there are two main types of advertisers in our dataset, one contains 46 advertisers and the other contains 54 advertisers. We report the means and standard deviations for each cluster in Table 6. In this table, state 1 represents the least engaged state and state 3 represents the most engaged state. Comparing the mean values of the clusters, we can see that most of the advertisers (86.66%) in cluster 1 already have a large proportion of very engaged consumers at the beginning of the sample period. In contrast, the advertisers in cluster 2 have lots of least engaged consumers (96.57%) at the beginning. The transition matrix of cluster 1 highlights that 86.6% of the consumers are likely to transfer to state 3 regardless of their current state. It indicates that advertisers in cluster 1 have engaging and effective ads. On the other hand, the transition matrix of cluster 2 shows the opposite pattern: consumers are likely to convert to the least engaged state no matter which state they are. This is reflected by the 96.69% of transition probability from any state to state 1. Therefore, the ads in this cluster are not very successful. The advertisers belonging to this cluster should think again about the ads content and design. The conversion probabilities also reveal consistent patterns. The average conversion probabilities for advertisers in cluster 1 are almost twice the average conversion probabilities for advertisers in cluster 2.

Table 6. Sample Size of 100 Advertisers

Probabilities	Cluster 1		Cluster 2	
	Mean	Sd	Mean	Sd
Initial_state1	0.0501	0.1491	0.9657	0.0326
Initial_state2	0.0824	0.2068	0.0206	0.0223
Initial_state3	0.8666	0.2801	0.0137	0.0115
Transition_from1_to1	0.0510	0.1483	0.9669	0.0324
Transition_from1_to2	0.0829	0.2067	0.0199	0.0223
Transition_ from1_to3	0.8660	0.2797	0.0131	0.0113
Transition_from2_to1	0.0511	0.1483	0.9670	0.0324
Transition_from2_to2	0.0828	0.2067	0.0198	0.0223
Transition_from2_to3	0.8661	0.2797	0.0131	0.0113
Transition_from3_to1	0.0511	0.1483	0.9669	0.0325
Transition_from3_to2	0.0829	0.2068	0.0199	0.0223
Transition_from3_to3	0.8660	0.2797	0.0131	0.0113
Conversion_state1	0.000214	0.000653	0.000115	0.000399
Conversion_state2	0.000215	0.000655	0.000117	0.000401
Conversion_state3	0.000218	0.000659	0.000119	0.000403
Number of Advertisers	46		54	

Interestingly, when we map the advertisers to the scatter plot of click-through rate versus conversion rate, the advertisers are not clustered (see Figure 5). Instead, advertisers in the first cluster are more widely distributed whereas the advertisers in the second cluster are more concentrated with mediocre conversion rates. The figure indicates the HMM results can reveal deeper patterns in the data that cannot be learned from summary statistics or descriptive analysis.

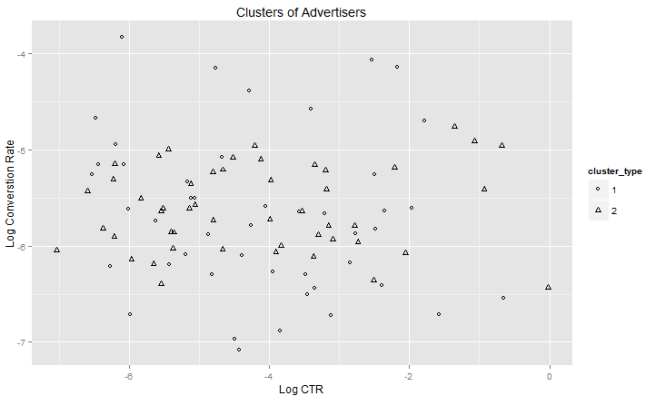

Figure 5. Location of Clusters on the Map of Click-Through Rate vs. Conversion Rate

Case Study of One Advertiser

In this subsection, we take one advertiser as an example and investigate more carefully into how the multi-screen media consumption is conducted. This advertiser has 1,502,860 impressions among which 25,755 impressions get clicked and 2,050 clicks are converted. The advertiser reached out to 467,172 consumers during out sample period. The average age of the consumers is 38.74 years old. 42.32% of the consumers are female. They own 1.48 devices on average. An average consumer conducted 2.17 clicks and 0.09 conversions during the sample period.

First, we report the transition parameters for this advertiser in Table 7. From left to right, each column reports the estimated coefficients corresponding to a pair of transition stages. For instance, the column βdi corresponds to the transition from disengaged state to interested state, and βei corresponds to the transition from engaged state to interested state. The HMM result is consistent with the model free evidence. From the negative coefficients of the mobile indicator and tablet indicator, we learn that advertisement on PC device is more engaging than that on Mobile or Tablet devices. From the positive mail indicator and homepage indicator, we learn that advertisement on the homepage and email context performs better than that on the rest types of the context. From the positive device stock coefficient, we can see that device diversity helps consumers to engage more. And finally, the negative coefficient of impression stock indicates that the marginal impact of an impression on conversion is declining over time.

Table 7. Transition Coefficients

	β_{di}	β_{ii}	β_{ei}	β_{de}	β_{ie}	β_{ee}
Mobile Indicator	-0.0070 (2.3e-5)	-0.0069 (4.0e-5)	-0.0068 (3.0e-5)	-0.0080 (2.5e-5)	-0.0080 (3.5e-5)	-0.0078 (2.6e-5)
Tablet Indicator	-0.0096 (2.3e-5)	-0.0094 (3.2e-5)	-0.0091 (2.4e-5)	-0.0110 (2.0e-5)	-0.0108 (2.7e-5)	-0.0104 (2.0e-5)
Mail Indicator	0.0073 (3.0e-5)	0.0078 (3.2e-5)	0.0048 (2.4e-5)	0.0084 (2.0e-5)	0.0089 (2.7e-5)	0.0051 (2.0e-5)
Homepage Indicator	0.0023 (2.3e-5)	0.0026 (4.2e-5)	0.0023 (3.0e-5)	0.0026 (2.6e-5)	0.0030 (3.5e-5)	0.0025 (2.6e-5)
Device Stock	0.0021 (2.4e-5)	0.0023 (3.2e-5)	0.0021 (2.4e-5)	0.0024 (2.0e-5)	0.0026 (2.7e-5)	0.0023 (2.0e-5)
Impression Stock	-0.0020 (2.3e-5)	-0.0024 (3.3e-5)	-0.0021 (2.4e-5)	-0.0023 (2.1e-5)	-0.0027 (2.8e-5)	-0.0024 (2.1e-5)
Click Stock	0.0086 (2.4e-5)	0.0099 (3.2e-5)	0.0088 (3.0e-5)	0.0097 (2.0e-5)	0.0113 (2.7e-5)	0.0100 (2.0e-5)
Intercept	-4.63 (2.9e-2)	-4.63 (3.2e-2)	-4.63 (4.0e-2)	-5.32 (3.5e-2)	-5.32 (5.6e-2)	-5.32 (3.8e-2)

Second, we simulate consumer cross-device media consumption activities. Specifically, we use the 324 sequences of all combinations of device in use for 4 impressions as the experimental setting. The model parameters are used to analyze how and when consumers convert. We observe the following: Of the people who see a mobile ad at the beginning of the campaigns, over 0.11% convert on the desktop within the next 3 impressions, but less than 0.06% convert on mobile if the impressions are always on mobile. Of the people who see a tablet ad before converting, about 0.09% convert on the desktop within the next 3 impressions, but less than 0.05% convert on the tablet if the impressions are always on the tablet. Of the people who see a desktop ad before converting, about 0.14% convert on the desktop as well within the next 3 impressions, and about 0.07% convert on other devices. This analysis shows that desktop is still the most important channel to the focal advertiser, especially if the consumers are in the later stages of the funnel. At the same time, other devices are also valuable, especially for unaware consumers that spend time on multiple channels for media consumption.

Contributions and Limitations

In this paper, we present a way to model how consumers respond to advertisement exposures on multiple devices. We quantify the marginal effect of one additional ad on conversion probability during the consumer journey to purchase. We allow the marginal effect to be device dependent, context dependent, and can change over time. Then we infer the initial and transition engagement level through the sequence of click and conversion activities.

The model is validated by a unique individual level dataset that contains 3.4 TB data. This Big Data provide great value for our analysis. First of all, clicks and conversions are relatively rare events compared to numerous impression occasions. Without a large amount of data, we cannot detect these rare events or draw conclusions with sufficient statistical power. Second, as summarized by Einav and Levin (2014), Big Data contain rich micro-level variation that can be used to identify novel behavior that is harder with smaller samples, fewer variables, and more aggregation. We leverage the variation in consumer demographics and ad exposures to evaluate the effectiveness of ads on conversion.

Using model free evidence and HMM estimation, we show the ad exposures cross devices are complementary to each other. Although the most effective device can vary from advertiser to advertiser, increasing device diversity consistently encourages consumers to move to more engaged levels. Using a random sample of 100 advertisers, we show that there are mainly two types of advertisers in our sample. One type has engaged consumers and effective ads delivered, whereas the other type has disengaged consumers and relatively ineffective ads. The type of the advertisers reveals the hidden structure of the data and cannot be learned directly from the descriptive analysis.

In terms of methodology, our study proposed and validated a distributed implementation of the non-stationary Hidden Markov Model. Using big data tools such as Hadoop, MapReduce, and Spark, we are able to handle hundreds of millions of records in structural analysis. Practitioners can potentially use this method on their real-world dataset in the day-to-day business.

There are several limitations of the current study that call for future work. First, our proposed work assumes each advertiser has its own conversion funnel. The conversion funnels are independent. Every consumer holds a position in each of these funnels. And her position in one funnel is not affected by the positions of other funnels. However, it is possible that the ads from multiple advertisers are competing for the consumer's attention and engagement. All else being equal, consumers that are exposed to more ads of competitive advertisers may be less likely to respond to the focal advertiser. On the other hand, it is also possible that there are advertising spillovers (e.g. Sahni 2013, Anderson and Simester 2013) because ads remind consumers of similar and non-exposed options. Therefore our analysis can be extended to allow interplay between advertisers. Second, we do not observe the advertising exposures beyond Company X's ecosystem. It is possible that the advertisers have reached out for consumers through many forms of ads on many platforms. Other researchers could combine the source of data from media platforms and advertisers to address this question. Third, our study fixes the number of states to be 3 for all advertisers. In reality, it is possible that consumers for different advertisers have different stages in the conversion funnel. In the future work of this study, we may use model fitness to choose the optimal number of states in the data. Finally, our study does not address targeted ads. The insight from out study may not be fully applicable to targeted ads.

Reference

Abhishek, V., Fader, P., Hosanagar, K. 2012. "Media Exposure though the Funnel: A Model of Multi-Stage Attribution," Working paper. Carnegie Mellon University.

Andrew, M., Luo, X., Fang, Z., Ghose, A. 2015. "Mobile Ad Effectiveness: Hyper-Contextual Targeting with Crowdedness," *Marketing Science* (35:2), pp. 218-233.

Baesens, Bart., Bapna, Ravi., Marsden, James., Vanthienen, Jan. Zhao, J.Leon. 2014. "Transformational Issues of Big Data and Analytics in networked Business", *MIS Quarterly* (38:2), pp. 629-631.

Business Insider, 2015. "Spending on Native Advertising is Soaring as Marketers and Digital Media Publishers Realize the Benefit", http://www.businessinsider.com/spending-on-native-ads-will-soar-as-publishers-and-advertisers-take-notice-2014-11.

Chen, Hsinchun., Chiang, Roger., Storey, Veda. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact". *MIS Quarterly* (36:4), pp. 1165-1188.

eMarket, 2014. "Marketers Expect Healthy Native Ad Spend Growth," http://www.emarketer.com/Article/Marketers-Expect-Healthy-Native-Ad-Spend-Growth/1011620.

Goldfarb, A. and Tucker, C. 2011. "Online Display Advertising: Targeting and Obtrusiveness," *Marketing Science* (30:3), pp. 389-404.

Goldfarb, A. and Tucker, C. 2011. "Search Engine Advertising: Channel Substitution When Pricing Ads to Context," *Management Science* (57:3), pp. 458-470.

Hauser, John., Urban, Glen., Liberali, Guilherme., Braun, Michael., 2009. "Website Morphing". *Marketing Science* (28:2), pp. 202-223.

Jordan, M. and Mitchell, T. 2015. "Machine Learning: Trends, Perspectives, and Prospects," *Science* (349:6245), pp. 255-260.

Li, B., Ghose, A., Liu, S. 2015. "Digitizing Offline Shopping Behavior Towards Mobile Marketing," In *Proceedings of the International Conference on Information Systems*, Dallas TX.

Li, H. and Kannan, P., 2014. "Attributing Conversion in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment," *Journal of Marketing Research* (51:1), pp. 40-56.

Microsoft, 2013. "How Marketers Can Drive Cross-screen Engagement," http://advertising.microsoft.com/en/cl/1932/cross-screen-research-report.

Montoya, R., Netzer, O., Jedidi, K. 2010. "Dynamic Allocation of Pharmaceutical Detailing and Sampling for Long-Term Profitability," *Marketing Science* (29:5), pp. 909-924.

Netzer, Oded., Lattin, James., Srinivasan, V., 2008. "A Hidden Markov Model of Customer Relationship Dynamics", *Marketing Science* (27:2), pp.185-204.

Rabiner, L., 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE* (77:2), pp. 257-286.

Randall, L., and Reiley, D. 2011. "Does Retail Advertising Work? Measuring the Effects of Advertising on Sales via a Controlled Experiment on Yahoo!" Working paper. Yahoo Research Labs.

Ryden, Tobias. 2008. "EM versus Markov Chain Monte Carlo for Estimation of Hidden Markov Models: A Computational Perspective," *Bayesian Analysis* (3:4), pp. 659-688.

Sahoo. N., Singh, P., Mukhopadhyay, T. 2012. "A Hidden Markov Model of Collaborative Filtering," *Management Information Systems Quarterly* (36:4), pp. 1329-1356.

Singh, P., Tan, Y., Youn, N. 2011. "A Hidden Markov Model of Developer Learning Dynamics in Open Source Software Projects," *Information Systems Research* (22:4), pp. 790-807.

Venture Beat, 2015. "Native Ads Are Replacing Banner Ads on Mobile – and Here's Why," http://venturebeat.com/2015/07/09/native-ads-are-replacing-banner-ads-on-mobile-and-heres-why/.

Xu, K., Chan, J., Ghose, A., Han, S. 2015. "Battle of the Channels: The Impact of Tablets on Digital Commerce," *Management Science*, forthcoming.

Xu, L., Duan, J., Whinston, A. 2014. "Path to Purchase: A Mutually Exciting Point Process Model for Online Advertising and Conversion," *Management Science* (60:6), pp. 1392-1412.